Volume 15 Number 4 1987 Nucleic Acids Research

Highly recurring sequence elements identified in eukaryotic DNAs by computer analysis are often homologous to regulatory sequences or protein binding sites

John W.Bodnar[1,*] and David C.Ward[1,2]

Yale University School of Medicine, Departments of [1]Human Genetics and [2]Molecular Biophysics and Biochemistry, 333 Cedar Street, New Haven, CT 06510, USA

Received September 3, 1986; Revised and Accepted January 12, 1987

ABSTRACT

We have used computer assisted dot matrix and oligonucleotide frequency analyses to identify highly recurring sequence elements of 7-11 base pairs in eukaryotic genes and viral DNAs. Such elements are found much more frequently than expected, often with an average spacing of a few hundred base pairs. Furthermore, the most abundant repetitive elements observed in the ovalbumin locus, the β-globin gene cluster, the metallothionein gene and the viral genomes of SV40, polyoma, Herpes simplex-1 and Mouse Mammary Tumor Virus were sequences shown previously to be protein binding sites or sequences important for regulating gene expression. These sequences were present in both exons and introns as well as promoter regions. These observations suggest that such sequences are often highly overrepresented within the specific gene segments with which they are associated. Computer analysis of other genetic units, including viral genomes and oncogenes, has identified a number of highly recurring sequence elements that could serve similar regulatory or protein-binding functions. A model for the role of such reiterated sequence elements in DNA organization and function is presented.

INTRODUCTION

DNA sequence elements which regulate gene expression and replication or participate in higher order chromatin structure, i.e., DNA-nuclear matrix interactions, are of significant current interest. Although a priori one might expect that only a single copy of a regulatory or structural sequence need be associated with each genetic unit, several lines of evidence suggest that such sequences may instead be highly reiterated throughout the entire genetic unit. For example, the proviral DNA of Mouse Mammary Tumor Virus (MMTV) has been shown to be under the regulation of glucocorticoid hormones and specific binding sites for the glucocorticoid receptor have been identified in the MMTV promoter region (1); however multiple copies of the receptor binding site sequence have also been found within the coding sequences of MMTV (2). Similarly, multiple copies of the progesterone receptor binding site sequence are found throughout the ovalbumin gene locus

(3,4). These observations are entirely consistent with the proposal of Yamamoto and Alberts in 1976 (5) that the regulation of gene expression by steroids would require multiple receptor-DNA interactions of varying affinities. Recently, Vogelstein and coworkers (6) have shown that the Drosophila actin gene is associated with the nuclear matrix at specific sites within a 3.5 kilobase region at its 5'-end and that the association appears to depend on multiple interactions, each of which alone is too weak to mediate attachment of the actin DNA to the nuclear matrix. Similarly, Mirkovitch et al. (7) have reported multiple sequence-specific attachment sites to the nuclear matrix within the histone gene cluster.

We reasoned that if a particular regulatory or protein-binding sequence element was actually repeated many times throughout a genetic unit then short oligonucleotides that make up that element would be found at higher than the expected random frequency. We have identified such highly reiterated oligonucleotide sequences by comparing a given sequence to itself by dot matrix analysis and by plotting the frequency distribution of all possible tetra-or pentanucleotides present within the genetic unit. The most abundant oligonucleotides observed were often overlapping sequences that could be aligned to yield 7-11 nucleotide long sequence elements. An analysis of several well studied genetic systems in which regulatory sequence elements have been characterized revealed that the computer generated sequence elements were identical or remarkably similar to regulatory sequences or protein-binding sites identified by prior biochemical or genetic studies. Analysis of other genetic units revealed additional highly recurring sequence elements, three of which, TNTTCTTT, GCCGCCGCCG and GGGCGGNG, were found in several different genes, including oncogenes. We propose that sites for DNA interaction with control factors are often sequences which are among the most prevalent within the gene segment under regulation. Furthermore, we suggest that one mode of gene regulation in eukaryotes by multiple sequence elements may be mediated by interactions between these sequence elements, or proteins that bind to them, and the nuclear matrix.

METHODS

To identify short DNA sequences that are highly overrepresented within a given segment of genetic information, we have used the DOT MATRIX and statistical distribution computer program, FINDSITES, available through the NIH PROPHET and GENBANK database systems.

Although dot matrix analysis normally has been used to look for sequence homologies between different DNA segments, here we have used the method to identify highly reiterated sequence elements within a DNA segment by comparing the sequence to itself. Successful identification of such elements depends on the length of match which is plotted. If the length is too short, i.e., 3 or 4 bases of homology, the plot is full of matches and consequently full of noise. If the length is too long, i.e., 7 or 8 bases, imperfect copies of a repetitive element are missed. By analyzing matches of 5 or 6 bases in length we can reduce the noise but still see imperfect matches quite well. This method of analysis, however, uses a large amount of computer time, so that long stretches of DNA can be done easily only if the investigator has a dedicated computer.

An alternative method to find highly recurring sequence elements utilizes oligonucleotide frequency analysis; this method uses less computer time and is amenable to the analysis of many kilobases of sequence at a time. In principle, a given pentanucleotide would be expected to be found in either orientation within a DNA sequence at a random frequency of once every 512 base pairs, assuming a base composition of 50% AT and 50% GC. If a particular sequence element were found twice every kilobase pair, the presence of that element alone would increase the frequency of any pentamers within it by 100 percent. By plotting a histogram for the frequency of all possible tetramers or pentamers in a given DNA sequence, we looked for oligomers that were found at higher than random frequency. This type of frequency analysis did not require that variations in overall base composition be normalized in any way. However, since the dinucleotide CpG is very much underrepresented in eukaryotic genomes (8,9), we were concerned that the oligomers containing CpG dinucleotides would skew the distribution of oligomer frequencies. Therefore, oligomers containing CpG were usually considered separately.

In general, when both methods of analysis were used on the same DNA sequence, the most prevalent oligomers found by the frequency analysis were "overlapping" oligomers which formed a part of the repetitive elements found by the dot matrix analysis. Oligonucleotide frequency analysis is the method of choice for identifying sequence elements that are dispersed throughout large blocks of sequence, while the dot matrix method is better for detecting elements clustered with a small segment of a larger sequence block.

RESULTS

Occurrence of highly recurring DNA sequence elements in several well studied eukaryotic DNAs.

To determine if highly recurring sequence elements might be prevalent in eukaryotic DNAs and if they truly might have some functional significance, we applied the techniques described above to several well studied eukaryotic DNAs. In this way we could identify such elements in these sequences by computer and then determine if they had functional significance by comparison with sequences that were already known to be protein binding sites or to have regulatory functions. Since the highly recurring sequence elements were identified without consideration of their orientation within a gene segment, the correlation was scored positive if the element was either homologous or complementary to a known regulatory signal.

a. Ovalbumin Gene Locus. About 20 kilobase pairs of the chick ovalbumin locus has been sequenced to date including the ovalbumin, X, and Y genes (4,10,11). A frequency analysis of these sequences is presented in Figure 1A. Only pentamers without CpG's are shown since the locus is highly AT rich and very few pentamers with CpG are present; the most prevalent (ACGTG) is found only 16 times in the 20 kb of DNA. In contrast, the most prevalent non CpG oligomers, ATTTT, TTTTT, TTTTC, TTTCT, TTCTT, TCTTT, and TGTTT, are each represented over 125 times.

The ovalbumin locus is under the control of progesterone (3,4), and the consensus sequence for the DNA binding site for the progesterone receptor has been determined (see Figure 1B). The 7 most prevalent pentamers found in this locus share homology with the progesterone receptor binding site. If we overlap the oligomers, 6 of the 7 pentamers form a 10mer (ATTTTTCTTT) which is a 9 of 10 match to the center of the consensus (Figure 2B), and the seventh (TGTTT) is also found in the progesterone receptor consensus. In all, a 9 of 10 match to the overlap 10mer (ATTTTTCTTT) is found once every 2300 bp throughout the ovalbumin locus; one would expect to find a 10mer with a single mismatch once every 13,100 bp at random. Interestingly, an 8 of 10 match to this consensus is found once every 80 bp in the ovalbumin, X and Y genes!

b. Metallothionein. The human and mouse metallothionein loci have been studied by several groups, and three separate upstream promoters elements have been identified (summarized in ref. 12). A frequency analysis of the human metallothionein II gene (13) is shown in Figure 2A. This demonstrates that 5 of the 6 most prevalent CpG containing pentamers (TGCGC, GCGCC,

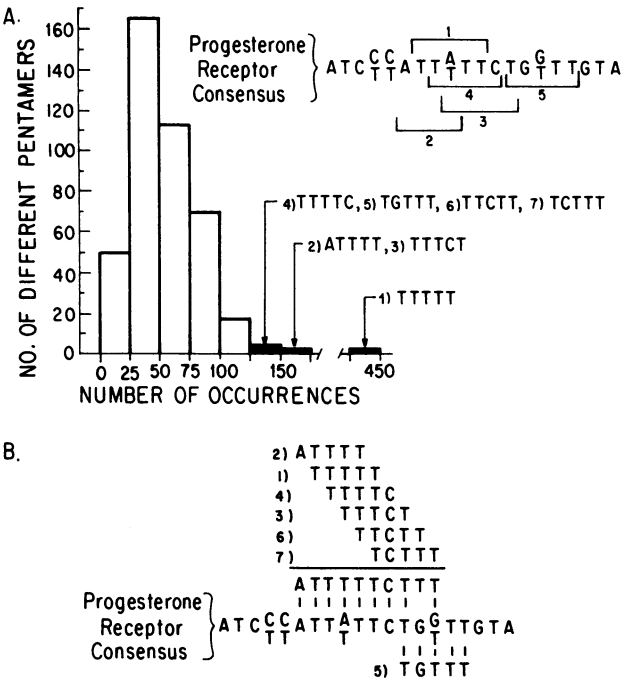

Figure 1. Pentanucleotide frequency analysis of the ovalbumin gene locus (Panel A) and the relationship of the most abundant pentamers to the progesterone receptor sequence consensus (Panel B).

CCCGG, CCGGC, and CGGCC) can be overlapped to form an 11mer TGCGCCCGGGCC. This 11mer is an exact match to the consensus sequence element implicated for control of expression of human metallothionein by heavy metals (12). This 11mer element occurs with $\leq$ 3 mismatches 16 times in the 1.7 kb metallothionein gene. However, in this case the elements are clustered so that 12 copies of the control element (8 of 11 match) are found in the 300 base pairs upstream of the mRNA cap site. A frequency analysis (data not shown) of the mouse metallothionein I gene (14) indicates that the most prevalent CpG pentamers in that gene also can be overlapped to yield an element homologous to the sequence required for control by heavy metals.

c. Herpes Simplex Virus, type 1. Approximately 15 kilobase pairs of HSV-1 DNA sequence is available on GENBANK. The frequency analysis of the pentamers in this sequence block is shown in Figure 2B. Due to the high GC content of HSV-1 the occurrence of CpG containing pentanucleotides is higher than in most DNAs. The most prevalent one percent of all the pentamers were

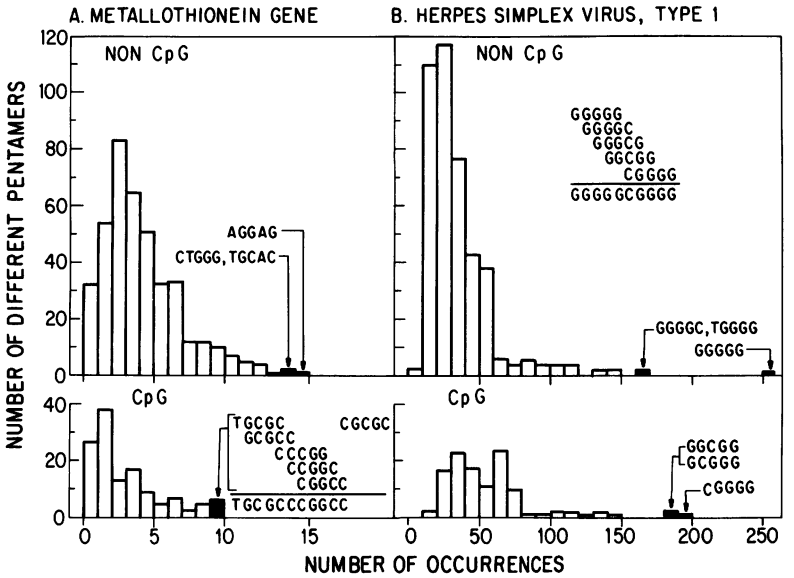

Figure 2. Pentanucleotide frequency analysis of the human metallothionein gene (Panel A) identifies six pentamers which overlap to form the sequence element TGCGCCCGGCC which is an exact match to the heavy metal control region (12) of the metallothionein gene. Panel B. Pentanucleotide frequency analysis of 15 kilobases of Herpes simplex virus, type 1, DNA. Panel A, pentamers without CpG; Panel B, pentamers with CpG. The most prevalent pentamers overlap to form the sequence element GGGGCGGGG.

the nucleotides GGGGG, GGGGC, GGGCG, GGCGG, CGGGG, and ACCCC. The first five of these are overlapping and make up the sequence element GGGGCGGGG (Figure 2B). This sequence element is virtually identical to the elements shown to be involved in control of transcription of both the HSV-1 thymidine kinase (TK) gene (15,16) and other HSV-1 "immediate early" genes (17). The SP1 transcription factor also binds to the promoter domains of these HSV-1 genes and stimulates transcription 25-fold (18). The SP1 recognition sequence, GGGGC GGGG, is strikingly similar to the abundant sequence element (G_4CG_4) identified by computer analysis.

d. SV40 and Polyomavirus Genomes The frequency analysis of SV40 and polyomavirus DNAs also indicated the presence of overabundant oligomers (date not shown). In SV40, the CpG pentamers which occur most frequently are GGGCG, GGCGG, GCGGA, and CGGAG. These are all overlapping oligomers which yield the sequence element GGGCGGAG. This sequence is found imperfectly repeated twice within the SV40 21 base pair repeats which have

been implicated in control of both transcription and DNA replication (19,20). The promoter-specific transcription factor SP1 also has been shown to interact specifically with the GGGCGG sequence motif within the 21 bp repeat (21,22).

The most prevalent non-CpG containing pentamers in SV40 DNA are TTAAA, TAAAA, AAAAA, and AAAAT which are overlapping in the sequence element TTAAAAAT. This element occurs 26 times in the SV40 genome with 1 or less nucleotide mismatches, and thus occurs with a frequency of once every 200 bp There is currently no data which suggests any biological function for this sequence element.

While the oligonucleotide frequency analysis of the entire SV40 genome failed to identify the three T-antigen-binding sites at the SV40 origin of DNA replication, the hexanucleotide sequence GGCCTC, representing the core contact sequence (23) was identified by dot matrix analysis of 600 bps surrounding the origin region (data not shown). Similarly, a dot matrix analysis of the polyoma replication origin region (24) identified the presence of the repeated oligomer AGAGGC. This is identical to the consensus for the polyoma large T-antigen binding site determined by Cowie and Kamen (25). In addition, a frequency analysis of the entire polyoma genome (data not shown) yielded the sequence AGAGGAGG from overlapping oligomers which is found with one mismatch or less once every 440 base pairs. The overlap oligomer contains a 5 of 6 match to the large T-antigen consensus suggesting that there might be multiple low affinity T-antigen binding sites throughout the polyoma genome.

e. Beta-globin. Over 20 kilobase pairs of the human β-globin locus are available on GENBANK. The frequency analysis of this locus, reveals that it contains three sets of overlapping oligomers that yield the sequences TTTTATTTTT, TTTCTTT and AAACGTG. Recently Plumb et al. (26) have defined an 18 nucleotide sequence in the 5' end of human β-globin gene which is protected from nuclease digestion by a protein factor (C4) isolated from chick erythrocytes. That sequence contains an exact match of one of the overlap oligomers (TTTCTTT) and a 8 of 10 match to the second (TTTATTTTT). Thus, overlapping of the most abundant pentamers in the entire β-globin locus yields a 16 of 18 nucleotide match to the C4 protein bind site. It should be noted, however, that no biological significance has been ascribed as yet to the C4 protein binding site or the AAACGTC sequence.

f. Mouse Mammary Tumor Virus (MMTV). The frequency analysis of the 8.4 kb MMTV proviral DNA identified three sets of overlapping oligomers (data not

TABLE I. Highly Recurring Sequence Elements in Viral Genomes

A. DNA Viruses	Sequence Element	Occurrence (≤ 1 mismatch)	Function-Sequence Relationship
Adeno-Associated Virus-2	CTCCACCA	1:310	-
	CGCAGAT	1:200	-
Adenovirus, type 2	GGGCGGG	1:160	a
	GCCGCCGCCG	1:640	b
	GGAGGAG	1:160*	c
	TGCAG(C/G)	1:430	-
Cauliflower Mosaic Virus	TTCTTCTT	1:140	f
Herpes Simplex, type 1	GGGGCGGGG	1:110	a
Minute Virus of Mice	AACCAACCA	1:230	d
Polyoma Virus	AGAGGAGG	1:440	c
Simian Virus-40	GGGCGGAG	1:880	a
	TTAAAAAT	1:200	-

B. Retroviruses	Element	Occurrence: Total genome	Occurrence: oncogene	Related Sequence
Avian Sarcoma Virus Y73	GAAGGAAG	1:210	1:160	e
Human T-cell Leukemia Virus-1 (HTLV-I)	CCGCCGCC	1:430	-	b
Fujinami Sarcoma Virus	AGCAGCAGC	1:430	1:190	-
Avian Myelocytomatosis (MC-29)	GCCGCCGCCG	1:380	1:160	b
Human T-cell Leukemia Virus-3 (HTLV-III)	T(T/C)TTCTTNT	1:190	-	f
	TTTCCTT	1:250	-	-
	TTAAAAAT	1:250	-	-
Mouse Mammary Tumor Virus (MMTV)	TTTCCTT	1:230	-	-
	TTTTCTTT	1:100*	-	f
	CTGCGG	1:820	-	-
Maloney Murine Leukemia Virus	CCCAGGG	1:160	-	-

No prominent recurring element was seen in the genomes of Hepatitis B virus, Bovine papilloma virus, type 1, human papovavirus BK, Rous Sarcoma virus PR, Simian Sarcoma virus, murine spleen focus-forming virus.
a = GGGCGGNG, the GC box sequence; b = GCCGCCGCCG, discussed in text; c = AGAGGC, the polyoma T-antigen binding site; d = potential terminal protein binding site; e = E1A enhancer core (A/C)GGAAGTGA(A/C), f = steroid receptor consensus (Table II).
* with no mismatches.

shown) that form the sequences (TTTTCTTTTT) (TTTCCTT) and (CTGCGG) . Transcription of MMTV DNA is modulated by glucocorticoids (1) and the glucocorticoid receptor protein has been shown to interact directly with the sequence TGTTCT (27). The first highly recurring sequence element contains

a 5 of 6 match to the glucocorticoid receptor binding sequence and it is found with a single mismatch on average once every one hundred base pairs throughout the entire MMTV genome. This element may thus represent the multiple low affinity receptor binding sites postulated by Yamamoto and Alberts (5) to be required for hormonal regulation during transcriptional activation. The second and third sequences, TTTCCTT and CTGCGG, have as yet not been shown to have a functional regulatory role in MMTV gene expression.

Identification of abundant sequence elements in other vertebrate and viral genes.

The striking ability of simple statistical methods to identify known regulatory sequences or protein binding sites in well characterized genetic units prompted us to search for highly recurring oligonucleotide elements in other genes and viral genomes. Some of this data is summarized in Table I.

Both dot matrix and oligonucleotide frequency analysis of the Avian Myelocytomatosis Virus MC-29 genome (28) indicated that the sequence element GCCGCCGCCG is highly overrepresented. Interestingly, it is found imperfectly repeated 8 times within the 700 base pairs near the 5' end of the V-myc gene but only 10 times, either exactly or with one mismatch, within the entire MC-29 sequence (i.e., once every 380 bases). This is much more than the expected random occurrence of 1 in 35,000, a particularly low value that reflects the presence of 3 CpG dinucleotides in this 10mer. Several other genes also contained GCCGCCGCCG in high copy number. These included: human c-myc (1:1010 bases with $\leq$ 1 mismatch), mouse c-myc (1:585), human T24 bladder oncogene (1:720), dog insulin (1:325), mouse keratin (1:100), mouse and rat 45S ribosomal RNA precursor genes (1:210), adenovirus type 2 (1:660) and the human T-cell leukemia virus, type I [HTLV-I] (1:430). Thus, this element is highly recurring in several different oncogenes and oncogenic viruses. While the GCCGCCGCCG sequences are clustered near the promoters of the myc and T24 oncogenes, they are distributed throughout the 35 kb of the Ad2 genome.

Table I, shows the recurring sequence elements present in 14 of the 20 fully sequenced eukaryotic viruses which were available on the GENBANK database. In Cauliflower Mosaic Virus [CAMV], for example, the sequence element TTCTTCTT occurs 56 times in 8024 base pairs; all but one of these elements share the same orientation. In most retroviruses, the sequence elements identified were localized mainly in or near the oncogenes carried by the viruses. However, in HTLV III the element, TCTTCTTTT is distributed throughout the genome and is in the same orientation 50 of its 52

TABLE II. Comparison of a class of highly recurring sequence elements (HRSE) with known protein binding sites

MMTV HRSE	TTTTCTTTTT
Gluococorticoid Receptor consensus	TGTTCT
Ovalbumin locus HRSE	ATTTTTCTTT
Progesterone Receptor consensus	ATC(C/T)(C/T)ATTTTTCTGGTTGA (A/T)
β-Globin locus HRSE	TTTCTTTT
C4 Binding site	ATATTTTTTTTCTTTT
HTLV III HRSE	TTTTCTTT (C)
Concensus	TNTTCTTT

occurrences. The Adenovirus 2 genome has several different highly recurring sequence elements. These were related to the elements found in other viral and cellular genes. For example, the element GCCGCCGCC is shared with MC-29 and HTLV-I, the element GGGCGGG is similar to that seen in SV40 and HSV-1, the element GGAGGAG is related to the element AGAGGAGG found in polyoma and the AGGAG sequence is identical to the most prevalent pentamer found in the human metallothionein gene. Furthermore, the abundant sequence elements in HTLV III and MMTV are related to the highly recurring sequence elements found in the ovalbumin and β-globin loci (Table II). These latter sequences may represent a class of binding sites for steroid hormone receptors. Indeed, recent evidence (29) indicates that the regulatory elements for different hormone receptors do share structural features as both glucocorticoid and progesterone receptor proteins bind to the same sites in two hormonally regulated promoters.

The observation that certain abundant sequence elements (e.g., TNTTCTTT, GCCGCCGCCG and GGGCGGNG) were present in many different genes suggests that there may be only a limited set of small highly recurring sequence elements in the DNA of eukaryotes. Futhermore, since one of these abundant elements, GGGCGCNG, is essentially identical to the binding site sequence of the transcription factor SP1, one can speculate that the other elements may also have similar biological significance.

DISCUSSION

We have described methods to identify highly recurring sequence elements (HRSEs)in DNA, and have shown that the majority of eukaryotic genes and viral genomes examined contain such elements. These elements have the following general characteristics: 1) they are the appropriate size for protein binding sites (i.e., 7 to 18 bp long); 2) they are found many more times than expected at random (often with an average spacing of less than 200 bp); and 3) they are generally distributed throughout the sequence in exons and introns as well as control regions. It should be noted, however that the simple frequency analysis method used here can miss sequences which have dyad-symmetry, especially where the central nucleotide(s) is undefined (N). One sequence with dyad-symmetry (CCC$^{A}_{T}$GGG) was detected in MMTV, although another, the 12 nucleotide binding site (TGGCANNNTGCCA) for the nuclear factor 1 protein (30), was missing in Ad2 DNA, even though this sequence is reiterated (i.e., a 9 of 12 match occurs 49 times in 36kb). Modifications of the basic computer program can, and are, being written to facilitate the detection of such sequences.

Seven of the 11 HRSEs identified in 7 well characterized gene segments were perfect or near-perfect matches to known regulatory sequence elements or protein binding sites. Only the TTAAAAAT sequence observed in SV40, the AAACGTG in the β-globin gene cluster, and the TTTCCTT and CTGCGG sequences in MMTV had not been previously shown to have defined biological functions. Considering that over 80 kb of total sequence was included in this analysis, the probability of this correlation occurring by chance is extremely small. These striking results suggest that sites for DNA interactions with protein or control factors are often among the most prevalent sequences within the regulated gene segment.

Why should a regulatory sequence be so abundant both within and around a eukaryotic gene? Although prokaryotic regulatory proteins, such as the repressor proteins for λ phage and lactose operon, bind specific DNA sequences with high affinity, they all also bind to other DNA sequences, albeit with substantially lower avidity. Lin and Riggs (31), after studying the thermodynamics of the lac repressor-DNA interactions, concluded that simple repressor-type regulation could not function efficiently in eukaryotic nuclei since in that environment the non-specific interactions with the non-operator DNA would totally mask operator-specific binding. Assuming that eukaryotic proteins are similar to their prokaryotic counterparts, the

specificity of DNA binding in a nuclear milieu would dictate that certain recognition sequences be highly repeated.

The data presented not only support the model for gene regulation by steroid receptors proposed by Yamamoto and Alberts (5), but also suggest that this is a general concept that can be extended to many other types of eukaryotic genes. This further indicates that selectivity in gene activation requires a large continuous segment of chromatin structure to be altered, and that such changes can be achieved when a regulatory protein occupies multiple sites in the same genetic region. The cooperative effect of binding at sites several hundred base pairs apart within a given genetic unit could thus modulate further steps in transcription or replication.

How might the modulation of chromatin structure and gene expression be affected? We suggest that the regulation may be mediated by interactions between these sequence elements, or specific proteins that bind to them, and the nuclear matrix. Several lines of experimental results are consistent with this interpretation. First, there is a rapidly expanding body of evidence indicating that the nuclear matrix of eukaryotic cells is a dynamic protein scaffolding upon which many cellular processes, including DNA replication and transcription, occur (32-37). Second, steroid receptor proteins for both estrogens and androgens are highly enriched (5-12 fold) in the nuclear matrix fraction but only in cells derived from their respective target tissues and only in response to an appropriate hormonal stimulus [reviewed by Barrack and Coffey (38)]. The glucocorticoid binding sequence also has been shown to function as a hormone dependent transcriptional enhancer element that selectively increases the efficiency of transcriptional initiation (39). Zaret and Yamamoto (40) have further demonstrated that both reversible and persistent changes in chromatin structure in the vicinity of the receptor binding sites occur during the activation of the enhancer element. We have shown here that the most abundant oligonucleotide element in the entire MMTV genome has a 5 of 6 match with the glucocorticoid receptor protein binding site while the most highly recurring sequence element in the ovalbumin gene locus has a 9 of 10 match to the sequence at the center of the putative progesterone receptor binding site. Together, these observations support the postulate that one function of the hormone-receptor complex is to serve as a bridge between the specific recurring sequence element in the gene and a component(s) in the nuclear matrix. If this were to be a common phenomenon, one would expect to see multiple, sequence-specific interactions between the nuclear matrix and other

individual genetic domains. Indeed, recent studies have demonstrated such interactions in Drosophilia within the actin gene family (6), the histone gene cluster (7) and a major heat shock protein gene (6).

Model for the role of DNA-nuclear matrix interactions in chromatin organization and expression.

We propose that there are two major classes of DNA-nuclear matrix associations: 1) stable DNA-protein interactions which delineate specific DNA domains within the genome and 2) dynamic DNA-protein interactions which are multiple, more labile associations involved in subnuclear localization and activation of gene expression. A schematic illustration of how these interactions could participate in chromatin structure and function is shown in Figure 3. This model consolidates and builds upon previous proposals relative to nuclear organization and gene function (6,7,11,31,32,41-44).

The DNA of both eukaryotic and prokaryotic cells is organized into discrete superhelical domains, with an average size of about 50 kilobases (reviewed in 34,42 and 45). These DNA domains are organized by stable association with the nuclear matrix (42), most likely mediated by proteins which are very tightly (perhaps covalently) bound to the DNA (45,46). These are the interactions which organize the DNA domains both on the metaphase chromosome scaffold (44) and the interphase nuclear matrix (7). The tightly bound proteins (TBPs) are indicated by the black circles in Figure 3. This type of long range organization would then allow the proper topology of the DNA for chromosome replication and segregation (41), and the subnuclear localization of replication complexes where they could more efficiently recognize the replication origins (37,47,48).

We suggest that superimposed on this type of long-range DNA organization is a second type of DNA-nuclear matrix interaction characterized by multiple, dynamic sites of DNA binding to the nuclear matrix throughout a given DNA domain. When a DNA domain is inactive (domain I in Figure 3), the sites of attachment to the nuclear matrix are the stable sites at the ends of the domain, and the DNA is packaged into chromatin which is "condensed" and inaccessible for expression. This inactive chromatin is associated predominantly with the nuclear lamina or peripheral nuclear matrix (49-54). When the DNA domain is activated (domain II in Figure 3), sites near the stable attachment sites can be recognized by cell specific factors, and their interaction changes the state of the entire domain (e.g., through demethylation, changes in supercoiling, etc.). Such alterations would then allow the DNA throughout the domain to be recognized by other factors at many specific

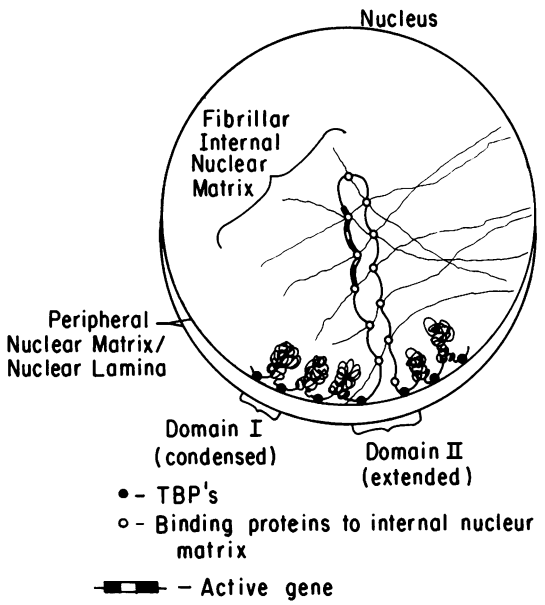

Figure 3. A schematic illustration of DNA-nuclear matrix interactions in chromatin organization and gene expression and the role of highly recurring sequence elements (O). See Discussion for details.

sites (i.e., the highly recurring sequence elements), and the DNA can then be extended into the internal nuclear matrix network by these multiple interactions (indicated by the open circles in Figure 3).

How can transcription or DNA replication proceed efficiently if the "extended" DNA domain is bound to the nuclear matrix at many sites? The plurality of binding sites does make it possible to read through, or dissociate, individual protein-DNA interactions without disrupting the overall structual organization of the "extended" DNA domain. Furthermore, recent studies on structure of the RNA polymerase III transcription factor A suggest a mechanism whereby individual proteins with multiple, independently folded, DNA-binding domains can remain bound to DNA during the passage of an RNA or DNA polymerase molecule (55). If such an "inch worm" mechanism for processing through preexisting protein-DNA complexes were operational at the level of DNA-matrix interactions, one could maintain both structural and functional requirements for gene expression or replication at optimal levels.

Is this model consistent with current concepts of chromatin structure

and organization? The division of DNA into "condensed" and "extended" forms corresponds extremely well to several other functional or morphological partitions of chromatin. Previous correlative studies (49-54,56-60) support the concept that "extended" DNA domains correspond to active chromatin, euchromatin, interior nuclear matrices, and early replicating DNA with HMG containing nucleosomes. In contrast, "condensed" DNA domains correspond to inactive chromatin, heterochromatin, peripheral nuclear matrix lamina, and late replicating DNA with H1 containing nucleosomes (49-54,56-60).

A final provocative feature of this model is that it provides a mechanism for efficient and specific expression of many different genes in different cell types with only a few regulatory proteins, a concept proposed by Gierer (61) and reviewed by Alberts (62) and Weintraub (56). If the transition from a "condensed" DNA domain to an transcriptionally active "extended" DNA domain occurs via a series of discrete steps (e.g., recognition of sites at domain ends, demethylation of domain DNA, matrix attachment at multiple intradomain sequences, promoter-matrix interaction, transcription-complex binding, transcriptional initiation) then a small number of different factors acting at each step in combinatorial fashion can specifically regulate the expression of many genes. Gene regulation by the mechanism proposed is costly for the cell, however, because an entire domain of DNA is utilized for each gene (or gene family) activated. It may be that eukaryotic organisms have evolved to pay that price in order to allow efficient regulation of many genes in many different cell types.

ACKNOWLEDGEMENTS

We thank Wayne Rindone, Harold Perry, and Bruce Stowe for assistance in using the PROPHET and GENBANK programs. We thank Alan Weiner, Sherman Weissman, Joan Steitz, Dan DiMaio, Maryellen Polvino-Rodnar, Helen Chao and Edith Gardiner for helpful discussion. This work was supported by NIH grants CA 16038, GM 32156 and AI 19973.

* Current address: Department of Biology, Northeastern University, 360 Huntington Avenue, Boston, MA 02115

REFERENCES

1. Payvar, F., DeFranco, D., Firestone, G.L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J.-A., and Yamamoto, K.R. (1983) Cell 35, 381-392.
2. Payvar, F., Firestone, G.L., Ross, S.R., Chandler, V.L., Wrange, O., Carlstedt-Duke, J., Gustafsson, J.-A., and Yamamoto, K.R. (1982) J. Cell. Biochem. 19, 241-247.

3. Mulvihill, E.R., LePennec, J.-P, and Chambon, P. (1982) Cell 24, 621-632.
4. Woo, S.L.C., Beattie, W.G., Catterall, J.F., Dugaiczyk, A., Staden, R., Brownlee, G.G., and O'Malley, B.W. (1981) Biochem. 20, 6437-6446.
5. Yamamoto, K.R. and Alberts, B.M. (1976) Ann. Rev. Biochem. 45, 721-746.
6. Small, D., Nelkin, B., and Vogelstein, B. (1985) Nucl. Acids Res. 13, 2413-2431.
7. Mirkovitch, J., Mirault, M.-E., and Laemmli, U.K. (1984) Cell 39, 223-232.
8. Nussinov, R. (1984) Nucl. Acids Res. 12, 1749-1763.
9. Subak-Sharpe, H., Burk, R.R., Crawford, L.V., Morrison, J.M., Hay, J., and Keir, H.M. (1967) Cold Spring Harbor Symp. Quant. Biol. 31, 737-748.
10. Heilig, R., Muraskowsky, R., Kloepfer, C., and Mandel, J.L. (1982) Nucl. Acids Res. 10, 4363-4382.
11. Lawson, G.M., Tsai, M.-J., and O'Malley, B.W. (1980) Biochem. 19, 4403-4411.
12. Schmidt, C.J., Jubier, M.F., and Hamer, D.H. (1985) J. Biol. Chem. 260, 7731-7737.
13. Karin, M. and Richards, R.I. (1982) Nature 299, 797-802.
14. Glanville, N., Durham, D.M., and Palmiter, R.D. (1981) Nature 292, 267-269.
15. McKnight, S.L. (1982) Cell 31, 355-365.
16. McKnight, S.L., Kingsbury, R.C., Spence, A., and Smith, M. (1984) Cell 37, 253-262.
17. Dynan, W.S. and Tjian, R. (1985) Nature 316, 774-778.
18. Jones, K.A. and Tjian, R. (1985) Nature 317, 179-182.
19. Hansen, U. and Sharp, P.A. (1983) EMBO J. 2, 2293-2303.
20. Bergsma, D.J., Olive, D.M., Hartzell, S.W., and Subramanian, K.N. (1982) Proc. Natl. Acad. Sci. USA 79, 381-385.
21. Dynan, W.S. and Tjian, R. (1983) Cell 35, 79-87.
22. Gidoni, D., Dynan, W.S., and Tjian, R. (1984) Nature 312, 409-413.
23. Jones, K.A., Myers, R.M. and Tjian, R. (1984) The EMBO J. 13, 3247-3255.
24. Soeda, E., Arrand, J.R., Smolar, N., Walsh, J.E., and Griffin, B.E. (1980) Nature 283, 445-453.
25. Cowie, A. and Kamen, R. (1984) J. Virol. 52, 750-760.
26. Plumb, M.A., Nicolas, R.H., Wright, C.A. and Goodwin, G.H. (1985) Nucl. Acids Res. 13, 4047-4065.
27. Scheidereit, C. and Beato, M. (1984) Proc. Natl. Acad. Sci. USA 81, 3029-3033.
28. Reddy, E.P., Reynolds, R.K., Watson, D.K., Schultz, R.A., Lautenberger, J., and Papas, T.S. (1983) Proc. Natl. Acad. Sci. USA 80, 2500-2504.
29. Von der Ahe, D., Janich, S., Scheidereit, C., Renkawitz, R., Schutz, G., and Beato, M. (1985) Nature 313, 706-709.
30. Nowock, J., Borgmeyer, U, Püschel, A.W., Rupp, R.A.W. and Sippel, A.E. (1985) Nucl. Acids Res. 13, 2045-2061.
31. Lin, S.Y. and Riggs, A.D. (1975) Cell 4, 107-111.
32. Berezney, R. (1979) The Cell Nucleus (ed. H. Busch) 1, 413-456. Academic Press, New York.
33. Ciejek, E.M., Tsai, M.J., and O'Malley, B. (1983) Nature 306, 607-609.
34. Hancock, R. and Boulikas, T. (1982) Int. Rev. of Cyt. 79, 165-214.
35. Hentzen, P.C., Pho, J.H., and Bekhor, I. (1984) Proc. Natl. Acad. Sci. USA 81, 304-307.
36. Jost, J.-P. and Seldran, M. (1984) EMBO J. 3, 2005-2008.

37. Pardoll, D.M., Vogelstein, B., and Coffey, D.S. (1980) Cell 19, 527-536.
38. Barrack, E.R. and Coffey, D.S. (1982) Rec. Prog. in Horm. Res. 38, 133-195.
39. Yamamoto, K.R. (1983) In Steroid Hormone Receptors: Structure and Function, Nobel Symposium 57; H. Erikkson, J.A. Gustafsson and B. Hogberg, eds. (Amsterdam:Elsevier/North Holland Biomedical Press) pp 285-306.
40. Zaret, K.S. and Yamamoto, K.R. (1985) Cell 38, 29-38.
41. Dingman, C.W. (1974) J. Theor. Biol. 43, 187-195.
42. Hancock, R. (1982) Biol. of the Cell 46, 105-121.
43. LaFond, R.E. and Woodcock, C.L.F. (1983) Exp. Cell Res. 147, 31-39.
44. Paulson, J.R. and Laemmli, U.K. (1977) Cell 12, 817-828.
45. Bodnar, J.W., Jones, C.J., Coombs, D.H., Pearson, G.D., and Ward, D.C. (1983) Mol. Cell. Biol. 3, 1567-1579.
46. Werner, D. and Petzelt, C. (1981) J. Mol. Biol. 150, 297-302.
47. Aelen, J.M.A., Opstelton, R.J.G., and Wanka, F. (1983) Nucl. Acids Res. 11, 1181-1195.
48. Vogelstein, B., Pardoll, D.M., and Coffey, D.S. (1980) Cell 22, 79-85.
49. Bouvier, D., Hubert, J., Seve, A.-P., and Bouteille, M. (1985) Exp. Cell Res. 156, 500-512.
50. Kaufmann, S.H. and Shaper, J.H. (1984) Exp. Cell Res. 155, 477-495.
51. Konstantinovic, M. and Sevaljevic, L. (1983) Biochim. and Biophys. Acta 762, 1-8.
52. LaFond, R.E., Woodcock, H., Woodcock, C.L.F., Kundahl, E.R., and Lucas, J.J. (1983) J. Cell. Biol. 96, 1815-1819.
53. Robinson, S.I., Nelkin, B.D., and Vogelstein, B. (1982) Cell 28, 99-106.
54. Setterfield, G., Hall, R., Bladon, T., Little, J., and Kaplan, J.G. (1983) J. Ultrastruct. Res. 82, 264-282.
55. Miller, J., Mc Lachlan, A.D., and Klug, A. (1985) EMBO J. 4, 1609-1614.
56. Weintraub, H. (1985) Cell 42, 705-711.
57. Brown, S.W. (1966) Science 151, 417-425.
58. Comings, D.E. and Okada, T.A. (1973) J. Mol. Biol. 75, 609-618.
59. Huberman, J.A. Tsai, A., and Deich, R.A. (1973) Nature 241, 32-36.
60. Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A., and Nag, A. (1984) Science 224, 686-692.
61. Gierer, A. (1973) Cold Spring Harbor Symp. Quant. Biol. 38, 951-961.
62. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. 1983) Molecular Biology of the Cell. Garland Publishing, Inc. N.Y. pp 444-445.